Journal of Mathematics and Statistics

Original Research Paper

Comparison of CART and CART ARCING to Classify Household Food Security of Tuberculosis Patients in Surabaya

[1]Iis Dewi Ratih, [1]Sri Pingit Wulandari, [2]Bambang Widjanarko Otok and [2]Yusnada Asa Nurani

[1]Department of Statistics Business, Faculty of Vocation,
[2]Department of Statistics, Faculty of Mathematics Computing and Data Sciense,
Institut Teknologi Sepuluh Nopember Surabaya, Kampus ITS Sukolilo, Surabaya, 60111, Indonesia

Article history
Received: 14-12-2017
Revised: 19-04-2018
Accepted: 25-05-2018

Corresponding Author:
Iis Dewi Ratih
Department of Statistics Business, Faculty of Vocation, Institut Teknologi Sepuluh Nopember Surabaya, Kampus ITS Sukolilo, Surabaya, 60111, Indonesia
Email: iisdewiratih@gmail.com

Abstract: Indonesia shall fulfill the residents need of food by its heterogeneous and big resource of food. But in fact, Indonesian food security is still far from expectation because of the maintenance of food resource in Indonesia has not been maximized yet. Meanwhile, households with tuberculosis patients in Surabaya tend to classify as food-insecure. This study conduct a classification of households with tuberculosis patients food security in Surabaya by using Classification and Regression Tree (CART) and Adaptive Resampling and Combining (ARCING). The purpose of this study is to know the influential factor of households food security and the accuracy of its classification. As the result, the way households obtain foods is the most influential factor of Indonesian food security in Surabaya. Other than that, this study shows that accuracy when using ARCING is higher than CART.

Keywords: ARCING, CART, Food Security, Tuberculosis

Introduction

Increasing number of inhabitants induces increasing need of food, home, energy and another things. Indonesia as developing country placed as fourth country with the biggest inhabitant in the world (DetikFinance, 2014). As the increasing of the inhabitants induces the better management of food resource so all the inhabitants may fulfill their need of food.

As far the management of food resource doesn't reach the maximum way so that Indonesia's food security is still far from what people have expected. Regarded to the Minister of Agriculture named Andi Amran said in Global Food Security, Indonesia ranked 71 from 113 country which meant good condition for Indonesia (Julianto, 2016). But this said is contrary from the reality. Moreover, based on the research conduct by (Wulandari *et al*., 2016) at least 56% households in Surabaya classified as undernourished while 44% households classified as nourished.

Food security contains some factors that correlate each other so to classify food security we should use nonparametric methods. One of nonparametric method we can use is Classification and Regression Trees (CART). CART is a classification method that done by decision trees technic. While Adaptive Resampling and Combining (ARCING) is an algorithm which can use to reduce the classification error. In this study, the use of CART and ARCING are to classify food security of households with tuberculosis patients in Surabaya and gain information about influential factors. The result of this study is hoped to help the government to develop the food security status in Surabaya and gaining new knowledge by using CART ARCING.

Materials and Methods

Data used in this study is a secondary data taken from another research conduct by (Wulandari *et al*., 2016). Data used is a food security data of households with tuberculosis patients in Surabaya (January-December 2016). The scooped of this data is 10 disctrict which are explained as Bubutan, Dukuh Pakis, Genteng, Gubeng, Sawahan, Simokerto, Sukomanunggal, Tambak-sari, Tegalsari and Wonokromo. There are one response variable and six predictors which explained in Table 1.

Table 1: Varible study

Variable	Information	Category
Y	Food security	1 = Nourished 2 = Undernourished
X_1	Availability of fundamental food (rice)	1 = Available ≥20 days each month 2 = Available <20 days each month
X_2	Distance between House-Market	1 = ≤ km 2 = > 2 km
X_3	Number of family member	1 = < 7 person 2 = ≥ 7 person
X_4	Education level of households' head	1 = Minimal elementary school 2 = Didn't go to school
X_5	Ways to get fundamental food	1 = Not indebted 2 = Indebted
X_6	Protein consumed	1 = Animal or Animal and Plant 2 = Plant or nothing

Stages of the data analysis are described as follows:

1. Describe the characateristic of food secutity of households in Surabaya
2. Forming Classification Trees (CART)
3. Classifying food security using CART ARCING
4. Comparing accuracy of CART and CART ARCING
5. Making decision

Classification and Regression Trees (CART)

Classification and Regression Trees (CART) is a classifying method done by using decision trees technic. Regarded to (Breiman *et al*., 1984), model yields based on the data scale of response variable, if the response variable is in scale of interval and ratio so the yielded model will be regression trees. And if the scale of response variable is in nominal and categorical data so the yielded model will be classification trees.

CART have some advantages which are (Lewis, 2000):

- Have no assumption requirement
- Easy to explore and to take decision even when the data used is complex and multivaraible
- The result of the analysis is simple and easily classify new data efficiently and it's also easy to interpret

But CART has a weakness which the result isn't stable or in other words a small changes in training data will change the prediction result of the yielded trees and give different result (Sutton, 2005).

Generally there are three steps in CART algorithm which is forming classification trees, pruning and selecting optimal classification trees.

Table 2: Splitter variation

Data scale	Ordinal	Nominal	Continuous
Number of sample	n	N	n
Level	L	L	-
Splitter variation	$L-1$	$2^{L-1}-1$	$n-1$

Forming Classification Trees

There are three different steps in forming classification trees. The first step is selecting the splitter. Every selection of the splitter is done based on one independent variabel. Table 2 shows how to select the splitter based on the data scale of independent variable where L is the number of category in each scale and n is number of data used in analysis.

The method of splitter selection is still in trial and error since there is no rule in selecting the best method (Lewis, 2000). There are some methods often used to select the best splitter which are Gini index, Infromation index, Twoing index and Entropy index. Gini index is often used as the selecting the best splitter method. Formula used to select the best splitter is written below:

$$I(t)=\sum_{i,j,=1} p(j|t)\,p(i|t), i\neq j \tag{1}$$

where, $I(t)$ is heterogeneity index of t node, $p(i|t)$ is proportion of class i in t node and $p(j|t)$ is proportion of class j in t node.

Selecting the best splitter to generate classification tree is done by using goodness of split criteria. Goodness of split criteria is a based value to select s splitter in t node. Goodness of split shows the decreasing of heterogeneity of a class with function written below:

$$\phi(s,t)=\Delta I(s,t)=I(t)-p_L I(t_L)-p_R I(t_R) \tag{2}$$

where, $\phi(s,t)$ is goodness of split value, $I(t)$ is heterogeneity function in t node, P_L is proportion of left node observation, P_R is proportion of right node observation, $I(t_L)$ is heterogeneity of left node and I(t_R) is proportion of right node.

Splitter with biggest value of goodness of split is the best splitter among the splitter since it has the capability of decreasing the heterogeneity at the most. Best splitter will show the important variable of classification tree. The best splitter will show the biggest variable score as this score showed the contribution of each variable in generating the classification tree. Below is the formula used to calculate the score to know the contribution of each variable in generating classification tree:

$$skor=\sum_{i=1}^{q}\phi(s,t_i) \tag{3}$$

where, $\phi(s, t_i)$ is goodness of split value in each node. This score yileds by adding each goodness of split value of every variable that act as surrogate for each node (q).

Second step of forming classification trees is determining terminal node. In this step we will determine if t node will be splitted or stopped as terminal node based on some citerias. A node will be terminal node if a node contain only one observation or reach minimum observation or if a classificaton tree reach certain level or depth.

Last step of CART analysis is class labelling for each terminal node. Class labelling is needed to characterize the classification result of each class based on response variable. Class labelling is done by using criteria below:

$$p\left(j_0 \mid t\right) = \max_j p\left(j \mid t\right) = \max_j \frac{N_j(t)}{N(t)} \tag{4}$$

where, $p(j|t)$ is proportion of class j in t node, $N_j(t)$ is the number of observation of j class in t node, dandan $N(t)$ is total number od observation in t terminal node, so the class label for t terminal node is j_0.

Another class labelling criterion in terminal node is using the classes of response variable which has the least observation. In case that in a terminal node each classes have the same number observation, class labelling will be based on the observation of the response variable which has the least observation (Steinberg and Golovya, 2007).

Pruning

The first yielded classification trees is a maximum classification tree which is big sized. This kind of trees is yielded based on the splitter selection rule and goodness of split criterion which usually stop by only one observation on each terminal node. The bigger the size of classification trees the bigger the risk of the estimator to exceed the true value (overfitting) or the estimator to deduct the true value (underfitting) (Lewis, 2000).

To avoid overfitting and underfitting, we need a good and optimum classification trees. To get this kind of trees we can do pruning. Pruning is done by deduct the node without decreasing the accuracy. So the size of the tree will not be too complex and the accuracy is still good. To know which tree is good to prune we can use cost complexity methods. This methods use complexity parameter noted by C, which its value increase as long as pruning process is happening. Formula used to calculate cost complexity function in T sub tree from maximum classification tree (T_{max}) with $C \geq 0$ is written below:

$$R_C(T) = R(T) + C\,|\tilde{\mathrm{T}}| \tag{5}$$

where, $R_c(T)$ is cost complexity measure or T tree complexity at C error value, $R(T)$ is resubtitution estimate or T tree error classification, C is complexity parameter or cost for each additional terminal node on T tree and $|\tilde{\mathrm{T}}|$ is terminal node on T tree.

Selecting Optimum Classification Trees

Selecting classification trees with too big size will cost bigger value of cost complexity. This may be happening because of the complexity of the data so we need to select an optimum tree which provide simple structure and small error. There are two methods we can use to select an optimum tree.

Test Sample Estimate

This methods used when the sample sized is too big. The first procedure to use this methods is dividing the sample into training and testing data where training data use to make form the tree and testing data use to estimate total error proportion of test sample estimate. Formula used to calculate test sample estimate methods is written below:

$$R^{ts}\left(T_t\right) = \frac{1}{N_2}\sum_{(y_n, j_n)\in L_2} X\left(d\left(y_n\right) \neq j_n\right) \tag{6}$$

where, $R^{ts}(T_t)$ is total error proportion of test sample estimate, N_2 is total number observation of training data and $X(d(x_n) \neq j_n)$ is statement said that 0 will be the value if the statement inside the bracket is false and 1 will be the value for the true. To estimate ther error proportion from the tree select an optimal classification tree (T_t) with $R^{ts}(T_t) = min\ R^{ts}(T_t)$ (Lewis, 2000).

V-fold Cross Validation Estimate

This methods is used when data sample is rather small. The procedure to use this methods is by dividing the number observation randomly to independently and same sized V fold. When the process of forming classification tree happens, one fold act as testing data while others act as training data. This procedure happen for V times with different division of data sample. Formula used to do V-fold cross validation estimate is written below:

$$R^{cv}\left(T_t\right) = \frac{1}{V}\sum_{v=1}^{V} R^{CV}\left(T_t^{(v)}\right) \tag{7}$$

where, $R^{cv}(T_t(v))$ is total error proportion of V-fold cross valdation estimate and V as the number of used fold. To estimate total error proportion from this methods we can select optimum tree (T_t) with $R^{cv}(T_t) = min\ R^{cv}(T_t(v))$ (Lewis, 2000).

Adaptive and Resampling Combining (ARCING)

Adaptive Resampling and Combining (ARCING) is a method which can reduce the error classification by reduce variance within group of classification. This methods is adapted from *Bootsrap Agregating* (BAGGING). In this methods we resample the data learning by certain probability. ARCING algorithm is written below (Breiman, 1998):

1. Take training data ($T^{(k)}$) from the first training data with probability as below:

$$p^{(k)}(l)=\frac{1}{N} \tag{8}$$

   where, N is total number of first training data.

2. Form classifier or splitter (C_k) from data $T^{(k)}$
3. Conduct training data (T) on C_k so it yields total classification error $m(l)$
4. Renew step -k+1 by another probability formula as written below for misclassified data:

$$p^{k+1}(l)=\frac{1+m(l)^4}{\sum\left(1+m(l)^4\right)} \tag{9}$$

5. Repeat step (1) to (4) as much as K, where K is the number of the replication
6. Combine K classifier (C_1,C_2,...,C_K) by *simple voting* to get optimum classification tree

Accuracy

There are some value to determine classification accuracy, one of them is *Total Accuracy Rate* (1-APER). *Total Accuracy Rate* (1- APER) is proportion of observation which predict as success. Function we used to calculate 1-APER is:

$$1-APER=\frac{total\ succes\ prediction}{total\ prediction} \tag{10}$$

Food Security

Food and Agriculture Organization (FAO) explain food security as four main aspect which are food availability, food accesibility, equal dispersion of foodm and food quality or food security (Pramita, 2016). Food availability can be measured by the availability of fundamental food in a month or a year. Food stablity can be measure by food availability and eating frequency of family member in a day. Based on food stability and food availability, we can gain food continuity of a households. Food quality can be measured by variation of consumed protein (Pramita, 2016).

Food accesibility can be measured by 3 measurement which are physic accessibility, social accessibility and economic accessibility. Physic accessibility is measured using distance between house to the market while social accessibility measured by the number of family member and education level of households' head and economic accessibility measured by households' way to get fundamental food (PKLIPI, 2009).

A households classified as nourished when it capables to fulfill their family member food needs in quantity and quality, secure, adequatem and attained. While undernourished household is a household that not capable to fulfill their family member food needs in certain time and in certain physicological standar of growth and health (Purwantini, 2014).

Results

Characteristics of Food Security Factors of Households with Tuberculosis Patients in Surabaya

Measurements of food availability in households refer to food availiblity and adequacy in quantity which fulfill household's consumptions in a month or a year period. Below is the percentage of fundamental food availibity based on the data.

Based on Fig. 1, we've known that 92 households of 124 households or about 74% have fundamental food or rice availibity which also means the availability of rice is at least 20 days or more in a month. While 32 other households or about 26% have fundamental food availibity in less than 20 days per month.

Measuring food stability can use the food adequacy and eating frequency of family member in a day. Third aspect in food security is food accesibility which can be measured by using the households' easiness of getting food. This easiness can be measured by physic accessibility, social accessibility and economic accessibility. Physic accessibility can be exlpained by nearest market location. While social accessibility can be explained by the number of familiy member and education level of households' head. Economic accessibility can be measured by ways to get fundamental food. Below is the characteristics of food accesibility of the households in Surabaya.

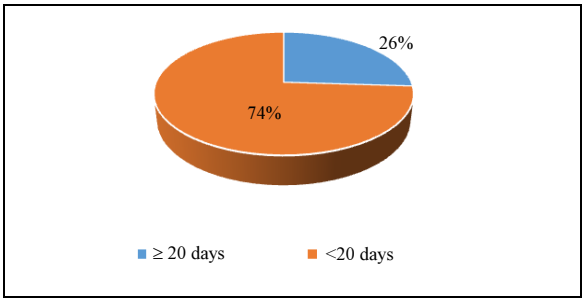

Fig. 1: Fundamental food availibity of households with tuberculosis patients in Surabaya

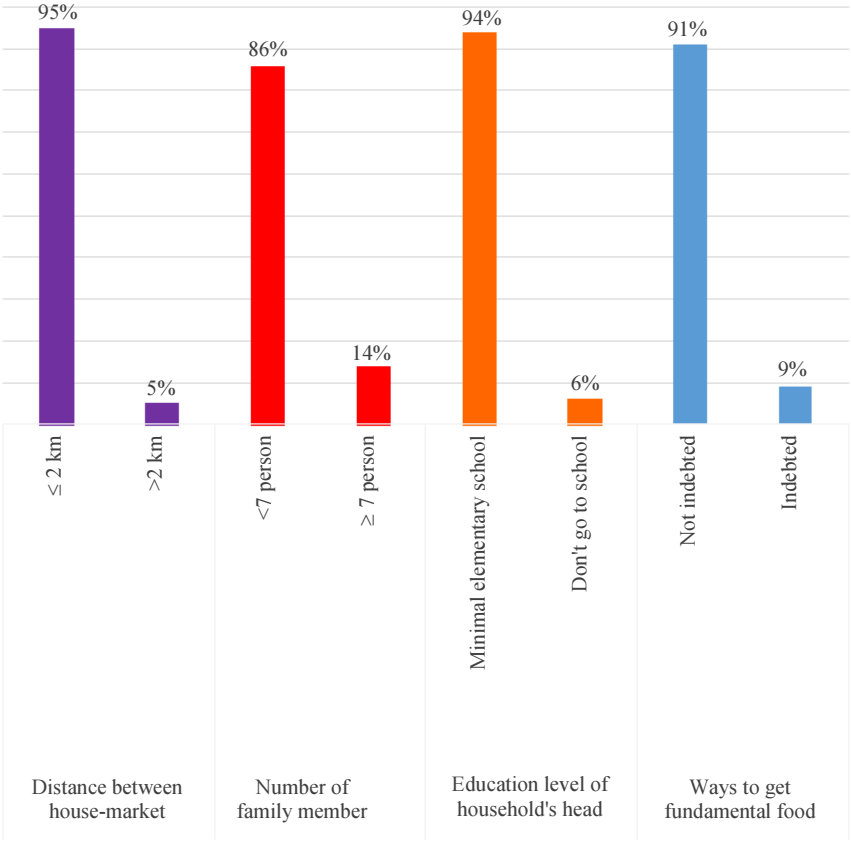

Fig. 2: Physical, social and economy access of the households with tuberculosis patients in Surabaya

Based on the data, 118 of 124 households or about 95% have less than 2 km in distance between house and market. While others or about 5% of total households have more than 2 km in distance between house and market. At least 107 of 124 households or about 86% have less than 7 person of family member while another 14% have more than 7 person of family member. 116 of 124 or about 94% (Fig. 2).

Households lived with household's head who has education level at least elementary school and another 6% of households liver with household's head who didn't go to school. Meanwhile at least 113 of 124 or 91% of households are get their fundamental food without indebted and another 9% get their fundamental food by indebted.

Last aspects of food security is food quality. To measure food quality we can use money spend on various side dish each day. Figure 3 define the variety of protein consumed by 124 households in this study. Based on Fig. 3, we've known that 90% or 112 households consumed animal protein or animal and plant protein while another 10% consumed plant protein or nothing.

CART Analysis of Households Food Security in Surabaya

Classification method used in this study is CART with selection splitter of Gini index and best splitter selection is goodnessof split. In this study, 124 observation are used where 99 observation are being training data and others as testing data. Based on training data, at least 54 households detected as nourished and 45 households as undernourished. Since the number of sample used in this study is relatively small we use 10-*folds cross validation estimate* to select optimum classification trees.

Splitter Possibilities

Variable act as splitter is selected based on splitter possibilites based on independent variable. Below is the splitter possibilities.

Based on Table 3, all the independent variables yield only one possibility of splitter.

Forming Maximum Classification Trees

First step of CART analysis is forming maximum classification trees. The very first idea of this step is selecting the importance splitter based on contribution score. This score actually show the contribution of each independent variables in generating maximum classification trees. Each score of independent variable shows in Table 4.

Based on Table 4, we've known that X_5 or way to get fundamental food scored 100,00. This also means that X_5 makes best reduce of the heterogeneity.

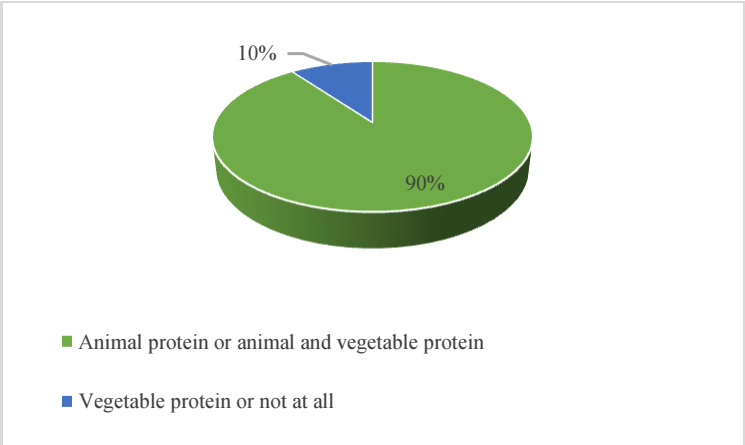

Fig. 3: Variety of protein consumed by households with tuberculosis patients in Surabaya

Table 3: Splitter possibilities

Var.	Variable name	Data Scale	Level	Splitter possibilities
X_1	Availability of fundamental food (rice)	Ordinal	2	2-1=1 splitter
X_2	Distance between House-Market	Ordinal	2	2−1=1 splitter
X_3	Number of family member	Ordinal	2	2−1=1 splitter
X_4	Education level of households' head	Ordinal	2	2−1=1 splitter
X_5	Ways to get fundamental food	Nominal	2	2^{2-1}−1=1 splitter
X_6	Protein consumed	Nominal	2	2^{2-1}−1=1 splitter

Table 4: Contribution score

Variable	Variable name	Score
X_5	Ways to get fundamental food	100,00
X_6	Protein consumed	90,61
X_4	Education level of households' head	84,07
X_3	Number of family member	30,25
X_2	Distance between House-Market	29,28
X_1	Availability of fundamental food (rice)	13,72

Way to get fundamental food can be measured using income. Associates with the limitation of this study, the income of households with tuberculosis patients will affect their way to get fundamental food since their income usually allocated to treat tuberculosis patients. Later, X_5 will be used as primary splitter or said as parent node.

After determining the parent node we can continue the splitting process with Gini index to form the maximum classification trees. Maximum classification trees is a tree with the most terminal node and depth. Based on this analysis we should continue to get optimum classification trees with 8 terminal node and depth of 5.

Pruning and Selecting Optimum Classification Trees

After forming maximum classification trees we shall see if the trees need pruning or else. Pruning main purpose is avoiding undefitting and overfitting. Pruning is done by using cost complexity minimum and 10-*folds cross validation estimate* so we get optimum classification trees. Pruning is yielded by using certain relative cost. Optimum classification trees is a tree with least relative cost. Figure 4 shows the relative cost plot of each pruning process.

Based on relative cost plot, we've known that optimum classification trees formed as the green line shows. The green line present at the last tree formed with 8 terminal nodes.

This optimum tree selection can also be supported by tree sequences tabulation in Table 5. Based on Table 5, first formed tree has 8 terminal node. Error cost of this tree is between 0,682 and 0,858 with complexity parameter about 0,000.

Since the first formed tree is optimum tree so it doesn't need pruning.

Optimum Classification Trees Analysis

The first formed tree is already an optimum tree since it has the least relative cost among other trees so it doesn't need pruning. The next step is optimum classification trees analysis which used to identify the accuracy of the classification. The optimum classification trees is shown in Fig. 5.

Table 5: Tree sequence

Tree number	Terminal nodes	Cross-validation relative cost	Resubtitution relative cost	Complexity parameter
1**	8	0,770±0,088	0,730	0,000
2	7	0,793±0,090	0,733	0,002
3	6	0,867±0,088	0,748	0,007
4	5	0,941±0,084	0,770	0,011
5	4	0,956±0,080	0,800	0,015
6	3	0,930±0,075	0,841	0,020
7	2	1,033±0,053	0,911	0,035
8	1	1,000±0,000	1,000	0,044

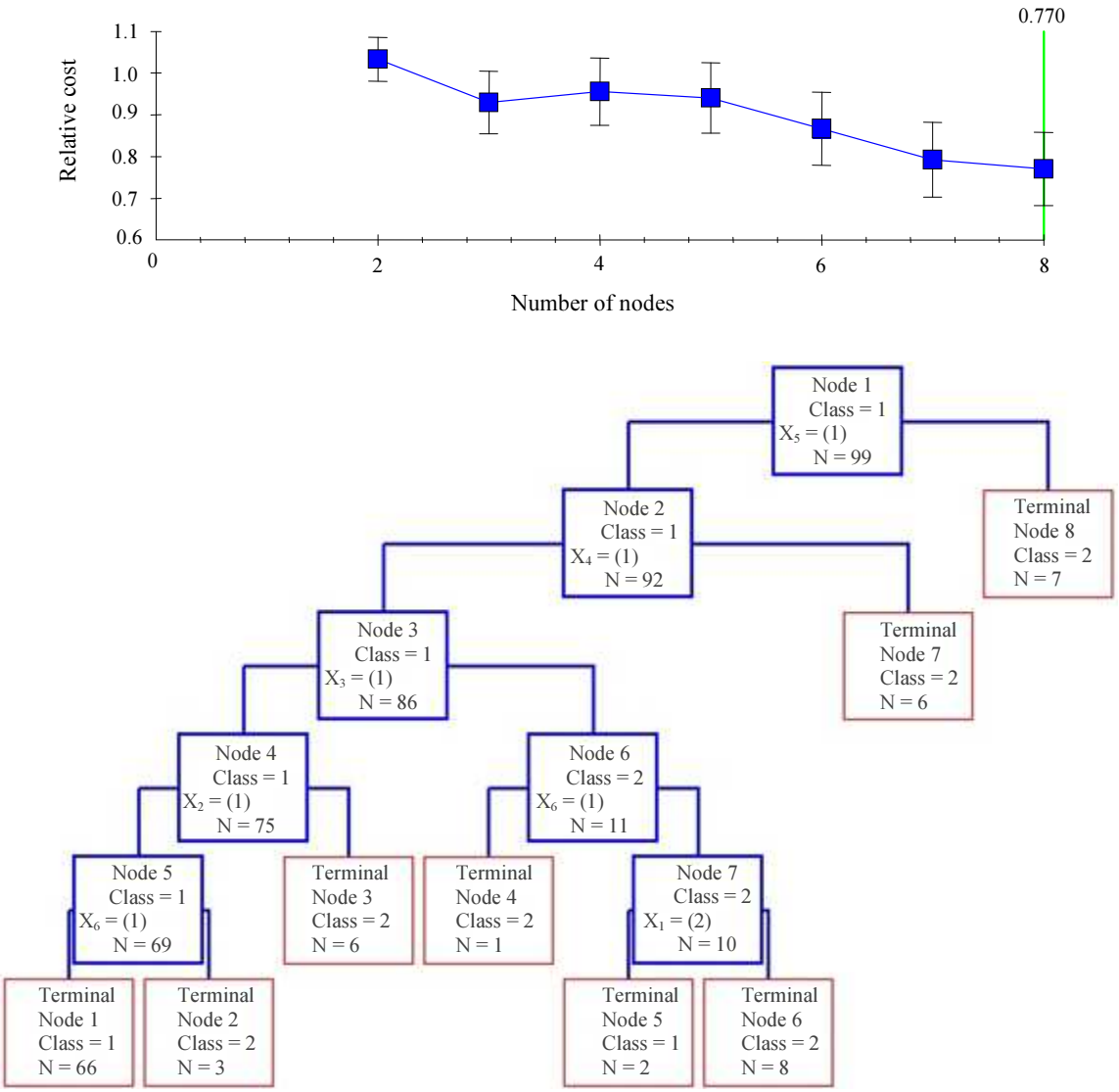

Fig. 5: Optimal classification tree

All training data split into two node which are left and right node based on way to get fundamental food. If a household get the fundamental food without indebted (category 1) so it will be grouped to left node and classify as nourished (class 1). While if a household get the fundamental food by indebted (category 2) so it will be grouped to left node and classify as undernourished (class 2). At least 92 households grouped into left node and 7 households grouped into right node. Splitting process is done recursively until it's impossible to split anymore and the tree formed terminal node.

Next step is class labeling to get the characteristics of each class based on the response variable. All of the terminal nodes yield has certain characteristics based on the predicted response variable. Table 6 give the information of number of terminal node and number of households grouped into each terminal node.

Based on Table 6 we've known that terminal node 1, 4 and 5 labeled as class 1 or nourished households. Meanwhile terminal nodes 2, 3, 6, 7 and 8 labeled as class 2 or undernourished households.

Sequential structure of terminal node 1 indicate that households that consumed animal protein, or plant and animal protein, distance between house-market less than 2 km, number of family member less than 7 person, lived with households' head who at least graduate from elementary school and without indebted in getting fundamental food. The number belongs to this class is 66 households.

At least 3 households belong to terminal node 2. Based on its sequential structure, households that belong to this grouped can be characterize as consumed plant protein or nothing, distance between house-market less than 2 km, number of familiy member less than 7 person, lived with households' head who at least graduate from elementary school and without indebted in getting fundamental food.

Based on terminal node 3 which classify as class 2 the characteristiccs of the households belong to this group are distance between house-market more than 2 km, number of familiy member less than 7 person, lived with households' head who at least graduate from elementary school and without indebted in getting fundamental food.

Terimnal node 4 classify as class 1 which characterize as consumed plant protein or nothing, number of familiy member more than 7 person, lived with households' head who at least graduate from elementary school and without indebted in getting fundamental food. There is only one household in this terminal node.

Terminal node 5 is also classify as class 1 which characterize as availability of fundamental food less than 20 days per month, consumed animal or animal and plant protein, number of familiy member more than 7 person, lived with households' head who at least graduate from elementary school and without indebted in getting fundamental food. There are two households in this terminal node.

Terminal node 6 is classify as class 2 which characterize as availability of fundamental food more than 20 days per month, consumed animal or animal and plant protein, number of familiy member more than 7 person, lived with households' head who at least graduate from elementary school and without indebted in getting fundamental food. There are 8 households in this terminal node.

Terminal node 7 is classify as class 2 which characterize as households who lived with households' head who didn't go to school and without indebted in getting fundamental food. There are 6 households in this terminal node.

Terminal node 8 is classify as class 2 which characterize as households who consumed plant protein or nothing at all. There are 8 households in this terminal node.

CART Analysis Accuracy

Based on Table 7 training data yields 59,6% which means 59,6% successfully predict in optimum classification trees. While testing data yields 52,0% for the accuracy. It's also means that the model accuracy reach 52,0%.

CART ARCING Analysis of Households Food Security in Surabaya

Table 8 shows the accuracy of CART ARCING analysis for the training and testing data. In this study we use 6 replication to generate CART ARCING analysis which are 25, 50, 75, 100, 125 and 150.

Based on Table 8 we've known that the accuracy for training data reach constant at value of 66,7% and testing data reach constant at the value 56%.

Comparation of Accuracy of CART and CART ARCING Analysis

Table 9 shows the information of accuracy of training testing data for households food security using CART and CART ARCING analysis.

Based Table 9 we've known that total accuracy rate yields from CART analysis is 59,6% for *training data* and 52,% for testing data. While total accuracy rate of CART ARCING analysis is 66,7% for training data and 56,0% for testing data. So we can conclude that by using CART ARCING this study increase the accuracy of the classification as much as 7,1% for training data and 4,0% for testing data.

Table 6: Terminal node class labeling

Class	Terminal nodes	Number of households
1 (Nourished)	1	66
	4	1
	5	2
2 (Under nourished)	2	3
	3	6
	6	8
	7	6
	8	7

Table 7: CART accuracy

		Prediction			
	Observation	Nourished	Undernourished	Total	Total accuracy rate
Training Data	Nourished	37	8	45	59,6%
	UNDERNOURISHED	32	22	54	
Total		69	30	99	
Testing Data	Nourished	8	2	10	52,0%

Table 8: Accuracy of CART arcing

	Trees combination					
Total accuracy rate (%)	25	50	75	100	125	150
Training data	66,7	50,0	66,7	66,7	66,7	66,7
Testing data	56,0	56,0	56,0	56,0	56,0	56,0

Table 9: Comparison of accuracy

	Methods						
		CART arcing					
Total accuracy rate (%)	CART	25	50	75	100	125	150
Training data	59,6	66,7	50,0	66,7	66,7	66,7	66,7
Testing data	52,0	56,0	56,0	56,0	56,0	56,0	56,0

Conclusion

Based on the analysis section we can conclude this:

1. Among 124 households that observed at least 74% have fundamental food availability which is rice for more than 20 days per month, 95% have less than 2 km of distance between house-market, 86% have less than 7 person of family member, 94% lived with households' head who at least graduate from elementary school and 90% consumed animal or animal and plant protein
2. CART analysis yield an optimum classification tree with 8 terminal node which 3 terminal nodes belong to class 1 or nourished households and 5 terminal nodes belong to class 2 or undernourished households. Importance variable in classification tree is way to get fundamental food with accuracy of 59,6% for training data and 52,0% for testing data.
3. Generating classification tree by CART ARCING has bigger accuracy which are 66,7% for training data and 56,0% for testing data

Things to recommend for the next study shall be the use of big data in analysis with the purpose of increasing the classification accuracy. Adding the study variable can be done to get more characteristics for each class. Other way, the next study can conduct by adding ensemble methods or another relevant methods. Furthermore for the government can take some concern for opening more job occupation or giving treatment asfor the tubeculosis patients so the income can increase and the number of the households who indebted when getting their fundamental food is decereased.

Acknowledgement

Thank you submitted to LPPM ITS who have supported this research in Laboratorium Research.

Author's Contributions

Iis Dewi Ratih: Participated in research and write the manuscript.

Sri Pingit Wulandari: Project Leader and Guide the first author and contributed to the writing of the manuscript.

Bambang Widjanarko Otok: Guide the first author and contributed to the writing of the manuscript.

Yusnada Asa Nurani: Collected and process data.

Ethics

The authors confirm that all the manuscript and there are no ethical issues involved.

References

Breiman, L., 1998. Arcing classifier. Ann. Statist., 26: 801-849.

Breiman, L., J. Friedman, C.J. Stone and R.A. Olshen, 1984. Classification and Regression Trees. 1st Edn., Chapman Hall, New York, London.

DetikFinance, 2014. Negara dengan Penduduk Terbanyak di Dunia, Indonesia masuk 4 Besar. DetikFinance, Maret.

Julianto, P.A., 2016. KOMPAS. KOMPAS Cyber Media (Kompas Gramedia Digital Group).

Lewis, R.J., 2000. An introduction to Classification and Regression Trees (CART) analysis. Annual Meeting of Society for Academic Emergency Medicine of San Fransisco, California.

PKLIPI, 2009. Ketahanan pangan rumah tangga di pedesaan: Konsep dan ukuran. ketahanan pangan dan kemiskinan dalam konteks de-mografi. Puslit Kependudukan-LIPI.

Pramita, H., 2016. Pemodelan penderita penyakit tuber-kulosis paru menurut stratifikasi ketahanan pangan ru-mah tangga di wilayah tengah kota surabaya dengan metode regresi logistik biner.

Purwantini, T.B., 2014. Pendekatan Rawan Pangan Dan Gizi: Besaran, Karakteristik, Dan Penyebabnya. 1st Edn., Pusat Sosial Ekonomi dan Kebijakan Pertanian, Bogor.

Steinberg, D. and M.D. Golovya, 2007. CART 6.0 User's Guide. 1st Edn., Salford Systems, San Diego.

Sutton, C.D., 2005. Classification and Regression Trees, Bagging and Boosting. Handbook Statist., 24: 303-329. DOI: 10.1016/S0169-7161(04)24011-1

Wulandari, S.P., D. Susilaningrum and I.N. Latra, 2016. Analisis ketahanan pangan rumah tangga terhadap kasus penderita penyakit tuberculosis dengan pende-katan geographically weighted poisson regression (Studi Kasus Pantai Pesisir Surabaya). Surabaya.